BIOINFORMATICS

Vol. 27 ISMB 2011, pages i266–i274
doi:10.1093/bioinformatics/btr241

Predicting site-specific human selective pressure using evolutionary signatures

Javad Sadri[1,2], Abdoulaye Banire Diallo[3] and Mathieu Blanchette[1,*]

[1]School of Computer Science, McGill University, 3630 University, Montreal, QC, Canada H3A 2B2, [2]Department of Computer Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran and [3]Department of Computer Science, Université du Québec à Montréal, Montreal, QC, Canada H3C 3P8

ABSTRACT

Motivation: The identification of non-coding functional regions of the human genome remains one of the main challenges of genomics. By observing how a given region evolved over time, one can detect signs of negative or positive selection hinting that the region may be functional. With the quickly increasing number of vertebrate genomes to compare with our own, this type of approach is set to become extremely powerful, provided the right analytical tools are available.

Results: A large number of approaches have been proposed to measure signs of past selective pressure, usually in the form of reduced mutation rate. Here, we propose a radically different approach to the detection of non-coding functional region: instead of measuring past evolutionary rates, we build a machine learning classifier to predict current substitution rates in human based on the inferred evolutionary events that affected the region during vertebrate evolution. We show that different types of evolutionary events, occurring along different branches of the phylogenetic tree, bring very different amounts of information. We propose a number of simple machine learning classifiers and show that a Support-Vector Machine (SVM) predictor clearly outperforms existing tools at predicting human non-coding functional sites. Comparison to external evidences of selection and regulatory function confirms that these SVM predictions are more accurate than those of other approaches.

Availability: The predictor and predictions made are available at http://www.mcb.mcgill.ca/~blanchem/sadri.

Contact: blanchem@mcb.mcgill.ca

1 INTRODUCTION

One of the central goals of comparative genomics is to use the comparison of genomes from different species to delineate functional regions in those genomes. Since most functional regions are under various degrees of negative selection, they tend to exhibit higher interspecies sequence conservation than their flanking neutral regions. Although mutations occur randomly in the genome (though at rates that may vary according to sequence context), the rate at which they become fixed in a population depends, among other things, on the fitness of the mutated individuals (Moran and Pierce, 1962). Whereas non-functional regions of the genome evolve mostly due to neutral drift (Kimura, 1983), functional regions are, for the most part, under negative selection, i.e. most mutations are deleterious. A consequence of this is that, over time, more mutations become fixed in non-functional regions than in functional regions. This principle is the foundation of phylogenetic footprinting, whereby one can hope to distinguish functional from non-functional regions of a given genome based on the observed number of mutations they have undergone during the evolution of a set of species. This approach has been used with success to identify all kinds of functional regions of the human and drosophila genomes (among others), including protein-coding genes (Dewey *et al.*, 2004; Gross and Brent, 2006), non-coding RNA genes (Dowell and Eddy, 2006; Pedersen *et al.*, 2006) and transcription factor binding sites (Loots and Ovcharenko, 2004; Moses *et al.*, 2004).

A number of generic approaches have been developed to identify sites that appear to be under negative selection based on comparative genomics. After identifying and aligning orthologous regions from two or more related species [e.g. using Multiz (Blanchette *et al.*, 2004a) or MLAGAN (Brudno *et al.*, 2003)], one can scan the alignment to identify regions where the sequence conservation is higher than expected. Early approaches dealt with a small number of genomes and evaluated conservation based on a sliding window strategy (Boffelli *et al.*, 2003; Frazer *et al.*, 2004; Margulies *et al.*, 2003). More recent approaches such as the very popular PhastCons approach (Siepel *et al.*, 2005) use a tree hidden-Markov model (HMM) to assign sites to one of several rate categories. A strength of this type of approach is that it can take advantage of the fact that most functional regions involve several consecutive sites, while avoiding the drawback of using a fixed-size window. This family of approaches has been used to identify likely functional regions in all kinds of species, including vertebrates (Margulies *et al.*, 2003, 2007; Thomas *et al.*, 2003), yeast (Kellis *et al.*, 2003) and drosophila (Stark *et al.*, 2007), among others, although most of these regions remain functionally uncharacterized to date.

With the number of genomes being sequenced quickly increasing, the prospect of accurately measuring evolutionary rates and selective pressure at *individual* sites became an achievable goal. Several approaches were developed for this purpose, including SCONE (Asthana *et al.*, 2007) and GERP (Cooper *et al.*, 2005), which both attempt to evaluate selective pressure on a site-by-site basis. What these approaches lose in specificity because of the fact that they do not combine signals from neighboring sites, they gain in sensitivity, since they are able to detect very small regions under selection.

A significant drawback of all these approaches is their assumption that the mutation rate in a given region or at a given site has not changed during the evolution of the set of species considered. Although negative selection in species ancestral to human and in sister species is clearly informative about selection in human, it is only an indirect predictor of it. Past results do not guarantee

*To whom correspondence should be addressed.

future performance in the stock market—neither does past selection necessarily implies current selection in human. There is considerable evidence that certain types of short functional regions such as transcription factor binding sites indeed turn-over quickly, causing a given locus to evolve neutrally in some species and under selection in another (Moses *et al.*, 2006). Recently, Siepel *et al*. introduced PhyloP (Pollard *et al.*, 2010), an impressive package that allows the detection of sites under negative or positive selection, while allowing changes in evolutionary rates over the branches of the phylogenetic tree (Siepel *et al.*, 2006). To our knowledge, this represents the best approach available to date for identifying individual sites under selection.

In this article, we propose a very different approach to the identification of functional sites in a given reference genome. Instead of *measuring* past selection at a site, we use the evolutionary history at a site (and at neighboring positions) to *predict* current (or very recent) selective pressure. More specifically, using whole-genome multiple alignments for a collection of 44 vertebrates (Margulies *et al.*, 2007; Miller *et al.*, 2007), we first reconstruct ancestral sequences (Blanchette *et al.*, 2004b; Diallo *et al.*, 2010) and identify evolutionary events at each site and along each branch of the tree. We then train machine learning classifiers to use the evolutionary history of a region to estimate the likelihood that it will have undergone a substitution in recent human history (since human–chimp divergence). Sites that are strongly predicted to remain conserved are likely functional sites in human. This approach has a number of key advantages. First, it does not require modeling or making assumptions about the evolution of functional regions of the genome process. Second, it is free to use whichever features of evolutionary history, and to weigh each feature as it wishes, in order to maximize the accuracy of the selective pressure prediction in human. A predictor may for example weigh more heavily conservation along primate lineages than along more distantly related lineages. It may also weigh more or less heavily conservation at neighboring sites.

This article is organized as follows. We first study how individual evolutionary events along specific branches of the phylogenetic tree relate to the probability of substitution in human. We then propose a set of machine learning classifiers, including a Naive Bayes classifier, a *k*-nearest neighbor classifier and a support-vector machine (SVM) classifier and show that some of them outperform recently proposed measures of sequence conservation. Finally, we show that the sites predicted by our best predictor (SVM) show better evidence of selection and function than those of existing conservation measures.

2 METHODS

2.1 Alignments and ancestral reconstructions

The genomes of 44 vertebrate species have been completely or partially sequenced to date. These include 7 primates, 25 other mammals, 2 marsupials, 3 birds and reptiles, 1 frog and 5 fish. Roughly half of these genomes are completely sequenced, while the other half is sequenced to at least 2× coverage (Margulies *et al.*, 2005). A whole genome 44-way multiple alignment was produced by Miller *et al*. (unpublished, based on the methology described in Miller *et al.*, 2007) using the MultiZ multiple alignment program (Blanchette *et al*., 2004a) and is available from the UCSC Genome Browser (Kuhn *et al.*, 2006). MultiZ produces a set of alignment blocks, where each block contains (presumably) orthologous regions from a subset or all the 44 species. When no reliable alignment can be found in a given species, that species is not included in the alignment block.

We applied a maximum likelihood ancestral sequence reconstruction approach (Diallo *et al.*, 2007, 2010) to infer, for each multiple alignment block, the ancestral sequence at each of the 43 internal nodes of the vertebrate phylogenetic tree. The program uses a tree-HMM approach to infer insertions and deletions over the branches of the tree, and then infers substitutions using a context-dependent substitution model. Blanchette *et al.* (2004b) and Diallo *et al.* (2007) have previously shown, using simulations, that the expected accuracy of this reconstruction can be as high as 99% base-by-base accuracy for early eutherian ancestors such as the Boreoeutheria ancestor, and that it is above 90% for almost all other ancestral nodes of the eutherian phylogeny.

2.2 Selection of unambiguously human conserved and mutated sites

A subset of human non-protein-coding genomic sites that could be unambiguously labeled as having undergone a substitution along the human branch since the human–chimp ancestry were identified. Several filters were applied to ensure that this set of sites is as enriched as possible for bona-fide mutated sites, and not the result of alignment errors. For site i to be considered as eligible, the following rules were applied:

- Site i must be conserved along the branches leading to the orangutan, gorilla and chimp.
- Sites $i-1$ or $i+1$ must be perfectly conserved between the human–chimp ancestor and both human and chimp.
- If site i is not conserved from the human–chimp ancestor to human, then it must be a transition (purine-to-purine or pyrimide-to-pyrimidine). Transversions are not considered, as they occur at different rates than transitions, which may introduce biases.
- If the human–chimp ancestral nucleotide is C, then it must not be followed by a G. This avoids CpG → TpG substitutions, whose elevated rate (Siepel and Haussler, 2004) may bias our analyses.

Finally, sites that satisfy all the above requirements are either called conserved or mutated, based on the event that took place between the human–chimp ancestor and human.

When applied to human chromosome 22, these filters resulted in the identification of ∼41 600 mutated non-coding sites, and roughly one hundred times more non-coding conserved sites. A subset of 41 600 conserved sites was then selected randomly from all conserved sites, to form a balanced set of 83 200 examples, which was divided into a 50 000-example training set and a 33 200-example test set.

2.3 Feature set definition and extraction

Various approaches were considered to encode the history matrix H into a set of features that can be used to train classifiers. Those that resulted in the best results were the following.

Feature Set 1: This contains the number of conservations and substitutions at each position within the window: $F_C^1(p)=\sum_b \mathbf{1}_{H(b,p)=C}$, $F_S^1(p)=\sum_b \mathbf{1}_{H(b,p)=S}$. This feature set thus contains $2(2w+1)$ integer features. Insertions and deletions are not explicitly accounted for in this feature set, but the presence of a large number of these events reduces the counts of conservations and substitutions and thus impacts the values of the features. An alternate feature set where all five types of events were counted separately was also evaluated but produced slightly worse results, probably because the increase in the size of the feature set is not counterbalanced by the informativeness of the new features.

Feature Set 2: This takes an orthogonal approach and counts events for each branch of the tree rather than for each site. Specifically, $F_C^2(b)=\sum_{p\in W(i,w)} \mathbf{1}_{H(b,p)=C}$, $F_S^2(b)=\sum_{p\in W(i,w)} \mathbf{1}_{H(b,p)=S}$. This feature set thus includes $85\times 2=170$ integer features, irrespective of window size.

A number of alternate feature sets have been considered including using the matrix H itself (or part of it corresponding to a smaller window around

the site). However, no predictor was able to take advantage of the richness of the provided information, probably in part because of the huge feature space involved.

2.3.1 Other measures of sequence conservation The PhyloP package (Pollard *et al.*, 2010) includes implementations of PhastCons (Siepel *et al.*, 2005), GERP (Cooper *et al.*, 2005), evolutionary rate likelihood-ratio test (LRT), as well as the PhyloP-SCORE. Each was run on the MultiZ 43-species multiple alignment obtained by removing the human sequence from the alignment, so that human conservation/mutations do not affect the scores produced. Default parameters were used for each algorithm. Each program outputs a score for each site in the dataset.

2.4 Classifiers

Our Naive Bayes approach was implemented in the straightforward manner, using uniform pseudocounts to estimate posterior probability features of the given class label. This classifier can be trained in a few seconds and make predictions equally quickly.

The k-nearest neighbor approach used the Euclidian distance between feature vectors to identify neighbors. We obtained best results on Feature Set 2, using $w=0$. Fairly large values of k (number of neighbors) produced better results; $k=400$ was used for the results reported here. This approach is substantially slower as the running time required to classify a single test example is proportional to the number of training examples. Nonetheless, it runs in a few minutes on a standard desktop computer.

We used the SVMlight package (Joachims, 1999) from a Matlab interface to train classifiers for Feature sets 1 and 2. While our efforts to reduce overfitting failed for Feature set 2 (which contains a large number of features), we obtained our best results on Feature Set 1, with window size $w=0$ and $w=1$. We obtained our best results using a radial basis function (RBF) with $\gamma=1.5$ (kernel parameter), $C=100$ (trade-off between training error and margin) and $J=0.7$ (relative cost of errors on positive examples). This is type of classification problem where the examples are fairly poorly separable results in a very large number of support vectors (more than half the training examples are retained as support vectors), which impacts running time and generalization. However, for Feature Set 1, the training and testing errors were essentially equal. Training on our 50 000 example took $\sim$1 h, and predicting on the 33 200 test examples took <10 min.

3 RESULTS

Our goal is to develop a machine learning predictor that will estimate the probability that a given site of the human–chimp ancestor will undergo a substitution along the branch leading to human, given the complete set of evolutionary events that took place in that region during vertebrate evolution (but excluding recent human evolution). Because the only way to predict positions where substitutions will become fixed is to predict their effect on fitness, sites that are predicted to remain conserved with high probability are likely to be functional ones.

It may seem counter–intuitive that in order to predict the selective pressure on a site in human, we purposefully ignore evolutionary events having taken place along that branch. In fact, although events along the human branch are excluded from our feature set, they play a very important role in our training set, as they form the label of each training example. We also underscore the fact that, although a substitution along the human branch is a very strong indicator of the absence of selective pressure at that site, such events are also extremely rare ($\sim$0.5% of sites). Thus, events along the human branch are much more productively used as labeled of a (artificially balanced) training set than as features.

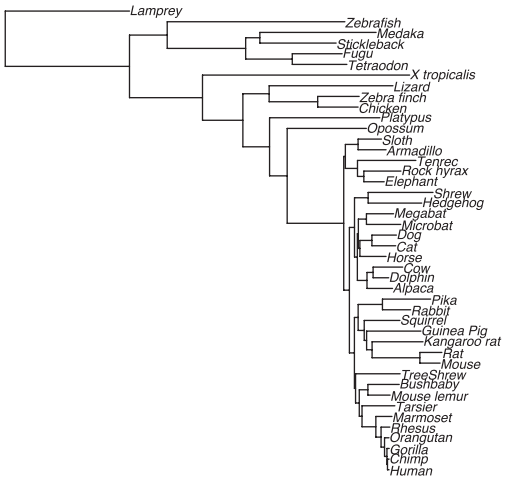

Fig. 1. Phylogenetic tree of the 44 vertebrate species considered to predict regions under selective pressure in human.

3.1 Training data

Our study is based on a balanced dataset consisting of all 41 600 non-coding sites from human chromosome 22 with unambiguous substitution along the human branch since the human–chimp ancestry and equally many non-coding sites with no substitutions, randomly sampled from the same chromosome. Coding regions were excluded from consideration because they can be accurately detected using a variety of approaches [see (Siepel, 2007) and references therein] and they obey fairly different rules than non-coding sites. A set of rules were applied to ensure that apparent human substitutions are not simply due to alignment errors or to increased substitution rate caused by sequence context effects (see Section 2). We then inferred the evolutionary history of each site, together with a 501 bp window centered on it, using a multiple sequence alignment of the genomes of 44 vertebrate species (Miller *et al.*, 2007) (Fig. 1) and a maximum likelihood ancestral sequence inference approach for both substitutions and indels (Diallo *et al.*, 2007, 2010) (see Section 2). The full history of a site i was then encoded as a matrix H_i with 85 columns (corresponding to the 85 branches of the phylogenetic tree, excluding that leading to human) and 501 rows (corresponding to the 501 human sites in the window surrounding site i), where the entry $H_i(\delta,b)$ for branch b at relative position $\delta\in\{-250,...,-1,0,1,...,250\}$ corresponds to the evolutionary event inferred on that branch at position $i+\delta$: C(onservation), S(ubstitution), I(nsertion), D(eletion) or G(ap). The 'Gap' event denotes the presence of a gap in both the ancestor and descendant of branch b. Our goal is to assess the extent to which the fate of site i along the human branch can be predicted from the matrix H_i.

3.2 Individual feature informativeness

We first measured how informative are individual events along each branch of the tree. This information can be measured by several means. First, we consider the question of whether the presence of orthologous bases in a given species (extant or ancestral) affects the likelihood of a conservation event along the human branch. A human site may have no detectable ortholog in a given

species s for several reasons: (i) Site i was inserted after the last common ancestor of s and human [denoted LCA(s,human)]; (ii) Site i was deleted since the LCA(s,human) along the lineage leading to s; (iii) Site i actually has an ortholog in s, but that and the surrounding sequence have diverged to the point where orthology cannot be detected (or, in the case of ancestral sequences, none of its descendant has a detectable ortholog). Figure 2a plots the likelihood ratio of human conservation in the presence or absence of an orthologous base on branch b at site $i+\delta$: $\log_2\left(\frac{\Pr[H_i(0,\text{human})=C|H_i(\delta,b)=C \vee H_i(\delta,b)=S]}{\Pr[H_i(0,\text{human})=C|H_i(\delta,b)=I \vee H_i(\delta,b)=D \vee H_i(\delta,b)=G]}\right)$. As expected, one observes that detectable orthology is relatively uninformative for primate species, as the vast majority of both functional and non-functional human sites have orthologs in these species. However, the value of orthology increases as we consider more divergent species, especially fast evolving ones such as rodents. This is because for highly diverged species, an increasing fraction of non-functional regions are either deleted or mutated beyond recognition, thus concentrating human functional sites in the fraction of sites with detectable orthologs. For example, sites with orthologs in other primate species are only ~7% more likely to be conserved than those without primate orthologs, but this number increases to 16% for other eutherians, 23% for marsupials and 33% for birds and reptiles. The trend presumably continues for more distant species such as fish, but we have insufficient data to observe it. It is interesting to consider how the events occurring at neighboring sites at position $i+\delta$ are also quite informative on the fate of site i, even for large δ. It appears that the presence of bases with a human ortholog even located 250 bp away from the current site is only marginally less informative than considering orthology at the site itself. This is due to the fact that functional regions and detectable orthology blocks are generally quite large.

Next, we ask whether the actual event (conservation or substitution) taking place at site $i+\delta$ along branch b brings any information on the fate of site i in human: $\log_2\left(\frac{\Pr[H_i(0,\text{human})=C|H_i(\delta,b)=C]}{\Pr[H_i(0,\text{human})=C|H_i(\delta,b)=S]}\right)$. Note that the probabilities at both the numerator and denominator assume that site $i+\delta$ has a detectable orthology between human and the species considered; they differ only in the fate of the site on branch b. Figure 2b reveals a trend that is quite different from that seen previously. Here, events taking place along branches close from human (e.g. primates) are the most informative. In fact, it appears that the main determinant of this likelihood ratio is how far away from the human lineage (defined as the set of branches leading from the root of the tree down to human) is the branch being considered. All branches along the human lineage have high log-ratios. This ratio decreases as we consider branches that are more distant from that lineage. It remains fairly large for ancestral branches close to it (e.g. those leading to lemurids and artiodactyls, for example), but decreases as we reach more distant branches (e.g. those leading to most extant non-primate species). This observation is new and quite significant. It suggests that as species diverge away from the human lineage, the function of some sites changes, making events along branches far from the human lineage less informative when it comes to studying the human genome.

Figure 2a and b fail to reveal an important feature of our data: detectably orthologous sites become increasingly rare as more and more distant species are considered, in such a way that <2% of human sites have orthologs in chicken, and this number drops to <0.1% in fish. A useful measure combining both the frequency of orthology and the informativeness of the evolutionary events taking place at or around the site is the mutual information between the event taking place along branch b at site $i+\delta$ and the presence or absence of conservation along the human branch:

$$I(H_i(\text{human},0);H_i(b,\delta)) = \sum_{\substack{a\in\{C,S,D,I,G\},\\ b\in\{C,S,D,I,G\}}} \left(\Pr[H_i(\text{human},0)=a,H_i(b,\delta)=b]\cdot \log\frac{\Pr[H_i(\text{human},0)=a,H_i(b,\delta)=b]}{\Pr[H_i(\text{human},0)=a]\cdot\Pr[H_i(b,\delta)=b]}\right).$$

Strikingly, the amount of mutual information is quite uniform over all the eutherian subtree. The single major exception are branches belonging to the old world monkeys subtree (the most closely related to human), which, despite their relative short length, bear significant amount of information over the fate of sites along the human branch, probably because their functionality is very similar to that in human. This would indicate that marks of selective pressure in old world monkeys is a very strong indicator of human selection, as previously suggested (Boffelli *et al.*, 2003). We also note that branches outside the eutherian subtree provide relatively little information on the fate of human sites, because the vast majority of human sites are aligned to gaps in those species.

3.3 Learning evolutionary signatures

We now turn to the problem of training a classifier to predict the fate of a human site based on the evolutionary history of its surroundings. Although one could in principle use the matrix H_i as the set of features from which to train a classifier, its size ($85\times501=42585$ features) is too large to hope to obtain good results with the size of the training set we are using (but see Section 4). Consequently, we considered two types of summary feature sets, parameterized by a window size parameter $w\leq250$, that reduce that number of features. Let $W(i,w)=\{i-w,...,i,...,i+w\}$ be the set of $2w+1$ sites surrounding i.

- Feature Set 1 contains $2\cdot(2w+1)$ features: for each position $j\in W(i,w)$, we record the total number of 'C' and 'S' observed at site j (summed over all branches).
- Feature Set 2 contains 2×85 features: for each branch b, we record the total number of 'C' and 'S' observed along branch b, summed over all sites in $W(i,w)$.

The decision to exclude counts of 'I', 'D' and 'G' events was made on the basis that these events are relatively rare in mammals and contribute little to the prediction problem while greatly increasing the number of features. Note, however, that although the counts of 'I', 'D' and 'G' events are not individually present as features, their presence is nonetheless reflected in the feature set through the counts of 'C' and 'S'. For example, in Feature Set 1, the total number of 'I', 'D' and 'G' events at a site is given by $85-n(C)-n(S)$.

3.4 The difficulty of the classification problem

We start by illustrating the difficulty of the classification task at hand, in order to calibrate our expectations. The fraction of human sites under selection has been estimated (based on human-mouse alignments) to be at least 5% (The International Mouse Genome

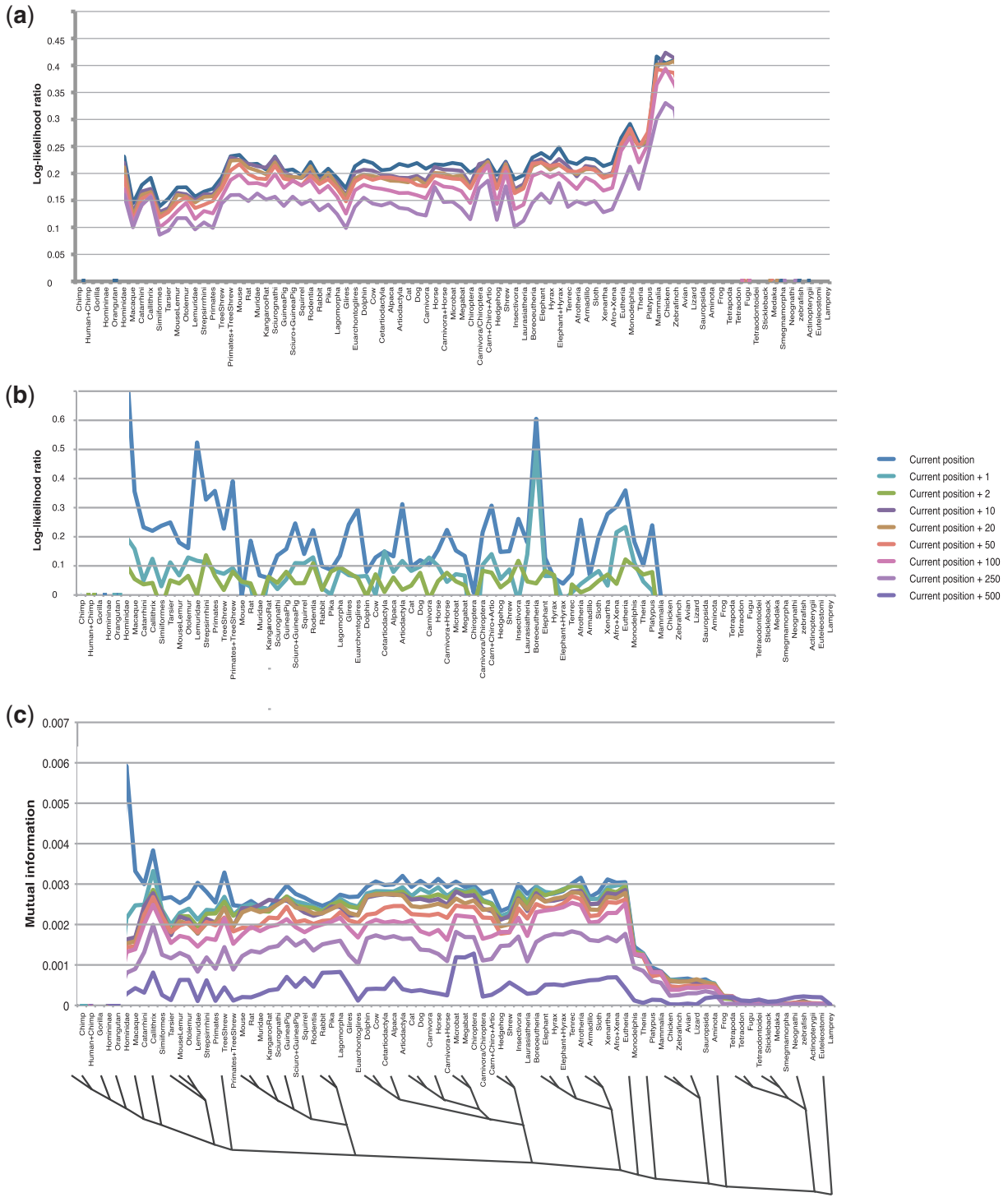

Fig. 2. Individual feature informativeness. (**a**) Log-likelihood ratio of human conservation in the presence or absence of orthologous bases in the given species and at the given offset from the considered position. (**b**) Log-likelihood ratio of human conservation in the presence of a conservation versus a substitution along the given branch and at the given offset. For both (a) and (b), ratios for non-mammalian species are too noisy and are not shown. (**c**) Mutual information between the presence of human conservation and the event along each branch at each offset.

Sequencing Consortium, 2002); more recent estimates place it between 4% and 7% (Margulies *et al.*, 2007). If we assume that the substitution rate in regions under selection is on average half the rate in neutral regions and that the probability of a neutral site mutating between the human–chimp ancestor and human is 0.5%, we obtain that $\frac{0.05\times0.0025}{0.95\times0.005+0.05\times0.0025}\sim2.56\%$ of mutated sites are under selection, while this fraction jumps to $\frac{0.05\times0.9975}{0.95\times0.995+0.05\times0.9975}\sim 5.01\%$ among conserved sites. Thus, both mutated and conserved sites are rich in non-functional sites, but conserved sites are almost two times richer in selected sites.

Consider, for the sake of example, a balanced training set of 1000 human-mutated sites and 1000 human-conserved sites. We expect the mutated sites to contain ∼975 neutral sites and 25 selected sites, while the conserved sites should contain ∼950 neutral sites and 50 selected sites. This has significant implications on the accuracy that our human-conservation predictor can be expected to obtain. For example, consider a classifier Ω that is able to distinguish selected from non-selected sites with 100% accuracy and that bases its prediction of human conservation prediction on this. The classifies Ω would predict 25+50=75 sites as conserved, of which 50 would be correct (67% positive predictive value). It would predict 975+950=1925 sites as mutated, of which 975 would be correct (50.64% negative predictive value). This is the best classification accuracy we can expect (under the assumption that the substitution rate at selected sites is half that at neutral sites). Remember, however, that our true goal is not to train a predictor to predict mutated versus conserved sites, but to predict non-functional versus functional sites. As we just saw, a relatively low accuracy at the mutated/conserved prediction task does not mean an equally low accuracy at the non-functional/functional prediction task.

3.5 Accuracy of predictors of human selection

Our set of 83 200 examples was divided into a training set of 50 000 examples and a test set of 33 200 examples, both with exactly the same number of human-conserved and human-mutated sites. The results reported in this section describe the performance on the test set, which was not used during the training phase.

Several types of classifiers were trained on each type of feature sets (see Section 2 for details on training and parameter tuning):

- *Naive Bayes*: classifiers calculate the posterior probability of the example label conditional on the set of observed features, under the (often unrealistic) assumption that features are independent. Despite their relative simplicity, Naive Bayes classifiers have proved remarkably useful in several settings (Rish, 2001).
- *k-nearest neighbor predictors (KNN)* (Shakhnarovish and Indyk, 2005): This simply obtain the probability of an example being positive based on a voting procedure using the k-nearest (most similar) training examples. This type of classifier can be very accurate for very large datasets, if an appropriate measure of similarity is available.
- *SVM* (Vapnik, 1995): This construct the optimal separating hyperplane between the two input classes to maximize the margin between the examples of the two classes. In the case of non-linearly separable classification problems, SVM uses kernel functions in order to implicitly map the feature vectors to a very high-dimensional feature space to obtain non-linear boundaries. After training SVM, support vectors are identified as those training examples that best define the boundary between the two classes. In the past few years, SVM have shown very good performance in a number of machine learning and pattern recognition problems (Cristianini and Shawe-Taylor, 2000).

All three types of predictors produced poor results when large values of w were used. The Naive Bayes and KNN classifiers performed best on feature set 2 (with $w=0$), whereas the SVM-based approaches were unable to handle the large number of features this set contains. On the other hand, the SVM approach was very effective at using the smaller number of features from Feature Set 1 and produced good results for both $w=0$ and 1. Figure 3 shows the positive predictive values (PPV, defined as the ratio of the number of true positive predictions to the number of positive predictions) obtained for each of these classifiers. Because we expect the fraction of functional sites in our balanced training and testing sets to be relatively small (probably around 5 to 10%), we only plot PPVs for prediction thresholds resulting in up to 20% of the test examples being predicted positive. The two SVM predictors (Feature Set 1, w=0 or 1) clearly outperforms all other approaches over much of the range of prediction threshold. The Naive Bayes and KNN predictors perform relatively poorly for high-confidence predictions, although they become competitive with the two SVM predictors at lower confidence calls.

In addition to these three types of classifiers trained on our two feature sets, we considered four previously proposed measures of sequence conservation, all implemented in the PhyloP package (Pollard *et al.*, 2010), that aim at detecting sites under selection (although not specifically along the human lineage):

- PhastCons (Siepel *et al.*, 2005): this approach identifies genomic regions with reduced mutation rate using a tree-HMM approach that assumes that neighboring sites are highly likely to have similar rates. It has been shown to perform very well at identifying various types of functional sites, in particular larger ones such as exons, enhancers and RNA genes.
- Genomic Evolutionary Rate Profiling (GERP) (Cooper *et al.*, 2005): this approach assigns a conservation score to each site in an alignment, independently of neighboring sites. It measures

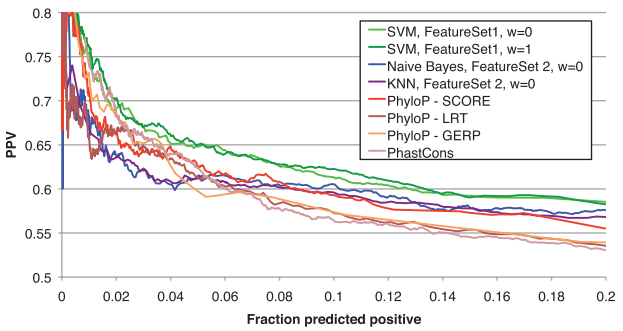

Fig. 3. Performance of various previously published measures of sequence conservation (PhastCons, PhyloP-SCORE, GERP, PhyloP-LRT), compared with predictors developed in this article. X-axis: fraction of test examples predicted as positive; Y-axis: positive predictive value (fraction of human-conservation predictions that are indeed human conserved).

the 'number of rejected substitutions', defined as the expected number of substitution per site minus the observed number.

- Likelihood-ratio test (LRT) (Margulies *et al.*, 2007; Pollard *et al.*, 2010): this approach assigns a *P*-value to the difference in the likelihoods of an observed alignment column under a null model of neutral evolution versus a model with an additional rate parameter that is estimated from the column.
- PhyloP-SCORE (Pollard *et al.*, 2010): similarly to the LRT, the SCORE test compares the hypotheses of neutrality to that of reduced or accelerated rate, but without the need to fit the rate parameter of the alternate hypothesis.

Each of these methods assigns a conservation score to each site i, based either on the evolutionary history of that site alone (in the case of GERP, LRT and SCORE) or based on the evolutionary history of site i and its surrounding sites (in the case of PhastCons). These conservation scores were calculated after excluding the human sequence from the alignment, to ensure that our class labels (event along human branch) do not taint our feature set.

The accuracy of the predictions made by these four measures of conservation are shown in Figure 3. PhastCons, the only measure that integrates signals from several consecutive positions, performs best for highly confident predictions (top 5% predicted as positive) but its accuracy decreases and quickly becomes worse than the other approaches at slightly more lenient thresholds. This is likely due to its excellent ability to detect relatively large regions under selection, but its inability to detect smaller ones or to identify weakly conserved sites within highly conserved regions. At more lenient thresholds (3–20% of test examples being called positive), PhyloP-SCORE outperforms the other three approaches. Note, however, that the two SVM-based predictors clearly outperform both PhastCons and PhyloP-SCORE over the full range of prediction threshold, often by fairly substantial margins.

3.6 Comparison of predictions

Figure 4 shows an example of the predictions made by our SVM predictor (FeatureSet 2, $w=1$), compared with PhyloP-SCORE and PhastCons. The most obvious difference is the increased resolution of the SVM and PhyloP approaches, which evaluate each base independently (though in the context of the two flanking bases in the case of SVM), compared with the HMM-based approach of PhastCons, which produces much smoother estimates. Whereas the PhastCons scores are extremely useful to define broad regions of conservation such as exons or enhancers, they are essentially useless when it comes to determining which bases within those regions are actually functional. For example, not all bases within an enhancer are equally constrained, because some are bound by a transcription factors and others not, and because not all bases within a TF binding site are equally important. Thus, PhyloP and SVM predictions provide an extremely valuable finer-grain estimation of selective pressure.

While it is difficult to compare the merits of each predictor based on a single example, Figure 5 provides a comprehensive comparison of the three sets of predictions, based on the ENCODE regions of human chromosome 22 (1.69 Mb in total) (ENCODE-Project-Consortium *et al.*, 2007). The top 10% of the sites predicted by each of the three methods was compared (this fraction is deliberately set slightly higher than the current estimates of the fraction of the genome under selection, to assess the ability of the methods to detect weak selection or selection on isolated sites). Clearly, the PhyloP and SVM approaches yield similar predictions (59% of the sites overlap), and quite different from the PhastCons sites (30–35% overlap only).

The predictions made by each method can be compared based on various external evidences of selection or function that were not part of their training. First, we compared the rate of human polymorphisms [HapMap project, phase III (International HapMap Consortium, 2007)] at sites predicted by each method or combination of methods. As expected, this rate is lowest for

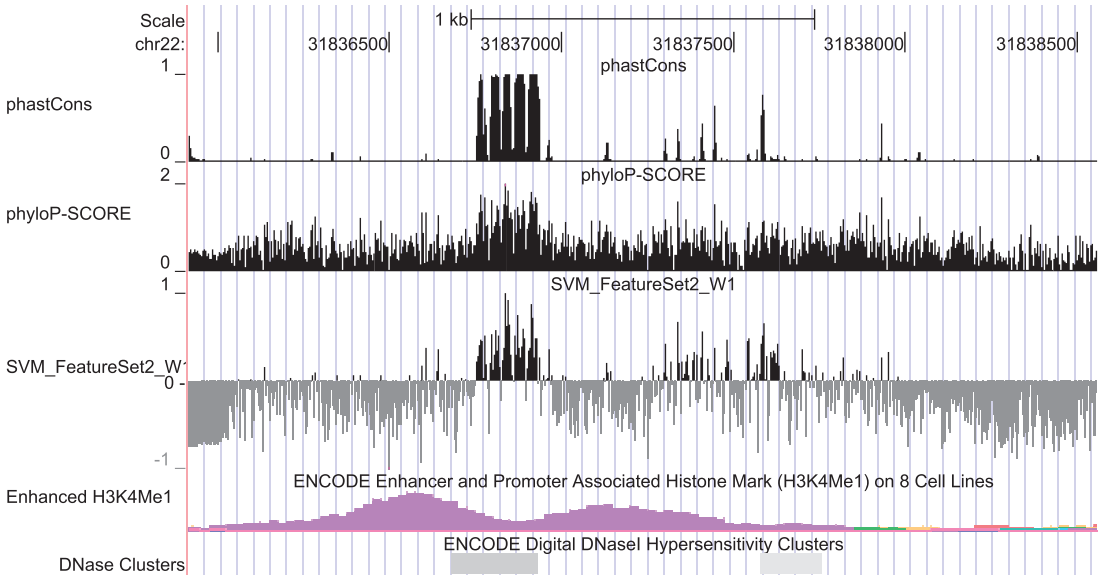

Fig. 4. Conservation scores obtained by three predictors: PhyloP-Score, PhastCons and our SVM predictor (featureSet 2, $w=1$) on human region chr22:31835800-31838550.

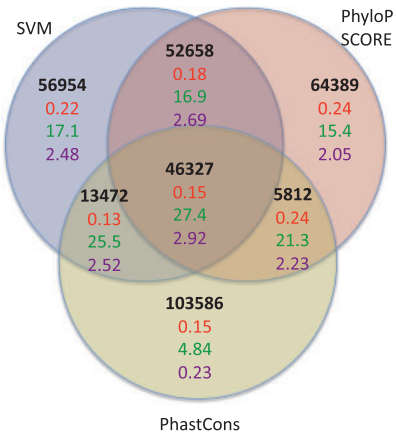

Fig. 5. Overlap of the top 10% most confident predictions made by three predictors on 1.69 Mb of human chromosome 22. Black numbers indicate the number of sites. Red numbers indicate the percentage of sites that are polymorphic in the human population, based on HapMap 3. Green numbers indicate the percentage of sites that overlap DNAse I hypersensitive regions in at least one ENCODE cell line. Purple numbers indicate the percentage of sites that overlap regions with H3K4me1 histone marks in one of eight ENCODE cell lines.

sites predicted by all three methods, suggesting that those are under the strongest selective pressure within the human population. Surprisingly though, the sites identified by PhastCons alone are equally depleted of polymorphisms, more so than those predicted by PhyloP-SCORE and SVM. The reason for this depletion, which contradicts the relatively poor positive predictive value of Phastcons compared with the other two predictors (Fig. 3), may lie in PhastCons's ability to identify fairly large regions under weak selection, such as non-coding RNA genes. We also note that the polymorphism rate is slightly lower in sites predicted by the SVM approach only, compared with those predicted by PhyloP, suggesting stronger selection on the former than the latter.

Second, we evaluated the evidence for a regulatory function of the predicted sites. Two types of experimental evidence were considered. Regions of the genome where the chromatin is open makes it possible for transcription factors to bind DNA. These regions can be identified by DNAseI digestion followed by hybridization or sequencing (Sabo *et al.*, 2006). DNAseI hypersensitive regions have been mapped in eight human cell lines as part of the ENCODE project (ENCODE-Project-Consortium *et al.*, 2007). As seen in Figure 5, 27% of the sites identified by all three predictors overlap one of these regions in at least one cell line. The numbers are lower for sites predicted by a single of the three methods, but still 17% SVM-only sites overlap DNAseI hypersensitive regions, a larger percentage than for PhyloP-SCORE-only sites or PhastCons-only sites. Another type of evidence of regulatory function is the presence of specific histone modifications, with H3K4me1 being strongly associated to enhancers (Heintzman *et al.*, 2009). This modification has also been mapped as part of the ENCODE project. Considering the union of H3K4me1 regions identified in eight cell lines as a set of likely enhancers, we observe again that sites predicted by all three methods overlap the most often with H3K4me1 regions, but that those predicted only by our SVM also overlap H3K4me1 marks significantly more often than those predicted only by one of the other two methods. Combined, these three types of evidence strongly suggest that the SVM predictor is better able to identify non-coding regulatory regions that are under selection in the human population.

4 DISCUSSION AND CONCLUSION

With the explosion of the number of vertebrates being sequenced comes the opportunity to detect marks of selection in subtler manners. Although a number of approaches have been proposed to identify regions that have evolved at a lower rate than the surrounding DNA, with some even allowing rates to change over the branches of the tree, we argue that it can be useful to consider the problem from a machine-learning perspective. Here, we show that it is possible to train simple classifiers to predict whether a human base is likely to mutate or not. The classification problem at hand is a challenging one, for several reasons: (i) the mutational process is a completely random one; only the rate of fixation varies between functional and non-functional sites. (ii) The fraction of functional sites is expected to be small. Because of these two properties, a predictor's ability to predict the label of a site (human mutated versus conserved) can only be achieved through the detection of function. Our conservation predictor is thus a predictor of functional sites.

Although we have restricted our analysis to sites from human chromosome 22, the next challenge will consist of using sites from the whole genome (∼100 times larger than our current dataset). A training set of that size (close to 10 million examples), combined with the fact that example labels are in large part random, poses significant challenges to most types of classifiers. Additional training examples could be obtained by considering substitutions along other branches of the mammalian tree (e.g. the branch leading to chimp), at the cost of losing some of the human specificity of our predictor. If these challenges are met, however, one can expect the development of highly accurate evolutionary signatures that may encompass not only the type of features used in this article, but also features such as the presence/absence of predicted transcription factor binding sites. It is difficult to imagine a more challenging yet fascinating classification problem!

Evolution has been conducting site-specific functionality assays for hundreds of millions of years. The ability to decipher the results of these experiments has and will continue to provide us with a wealth of information about our genome and the impact of mutations therein.

ACKNOWLEDGEMENTS

We thank Tomi Pastinen for useful discussions and the reviewers for their excellent suggestions.

Funding: Genome Canada and NSERC grants (to M.B.); FQRNT scholarship (to J.S.).

Conflict of Interest: none declared.

REFERENCES

Asthana,S. *et al.* (2007) Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.*, **3**, e254.

Blanchette,M. *et al.* (2004a) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Blanchette,M. *et al*. (2004b) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.*, **14**, 2412–2423.

Boffelli,D. *et al*. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Brudno,M. *et al*. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.

Cooper,G. *et al*. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–910.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, MA.

Dewey,C. *et al*. (2004) Accurate identiþcation of novel human genes through simultaneous gene prediction in human,mouse, and rat. *Genome Res.*, **14**, 661–666.

Diallo,A. *et al*. (2007) Exact and heuristic algorithms for the indel maximum likelihood problem. *J. Comput. Biol.*, **14**, 446–461.

Diallo,A. *et al*. (2010) Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, **26**, 130–131.

Dowell,R. and Eddy,S. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 40.

ENCODE-Project-Consortium, Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, **447**, 799–816.

Frazer,K. *et al*. (2004) Vista: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, 273–279.

Gross,S. and Brent,M. (2006) Using multiple alignments to improve gene prediction. *J. Comput. Biol.*, **13**, 379–379.

Heintzman,N. *et al*. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B. *et al*. (eds) *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, Cambridge, MA.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, MA.

Kuhn,R. *et al*. (2006) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D667.

Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, 217–221.

Margulies,E.H. *et al*. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res.*, **13**, 2507–2518.

Margulies,E.H. *et al*. (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA*, **102**, 4795–4800.

Margulies,E. *et al*. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, **17**, 760–774.

Miller,W. *et al*. (2007) 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res.*, **17**, 1797–1808.

Moran,P. and Pierce,A. (1962) *The Statistical Processes of Evolutionary Theory*. Oxford, Clarendon Press.

Moses,A. *et al*. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding sitespeciþc evolutionary model. *Genome Biol.*, **5**, R9.

Moses,A. *et al*. (2006) Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput. Biol.*, **2**, e130.

Pedersen,J. *et al*. (2006) Identiþcation and classiþcation of conserved RNA secondary structures in the human genome. *PLOS Computat. Biol.*, **2**, e3.

Pollard,K. *et al*. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Rish,I. (2001) An empirical study of the naive bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, Seattle.

Sabo,P. *et al*. (2006) Genome-scale mapping of dnase i sensitivity in vivo using tiling dna microarrays. *Nat. Methods*, **3**, 511–518.

Shakhnarovish,D. and Indyk,P. (2005) *Nearest-Neighbor Methods in Learning and Vision*. MIT Press, Cambridge, MA.

Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.

Siepel,A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Siepel,A. *et al*. (2006) New methods for detecting lineage-specific selection. In *Proceedings of the 10th International Conference on Research in Computational Molecular Biology*, Venice, Italy, pp. 190–205.

Siepel,A. (2007) Targeted discovery of novel human exons by comparative genomics. *Genome Res.*, **17**, 1763–1773.

Stark,A. *et al*. (2007) Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.

The International Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Thomas,J.W. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory.* Springer, New York, USA.